

1 16. (AMENDED) A method for filtering search results to
2 remove near-duplicates, the method comprising:
3 a) for each of a predetermined number of candidate
4 search results, determining whether the candidate
5 search result is a near-duplicate of another candidate
6 search result; and
7 b) if it is determined that the candidate search
8 result is a near-duplicate of another candidate search
9 result, then rejecting the candidate search result
10 wherein the act of determining whether a
11 candidate search result is a near-duplicate of another
12 candidate search result includes
13 i) comparing a cluster identifier of the
14 candidate search result with that of the other
15 candidate search result, and
16 ii) if the cluster identifiers of the two
17 candidate search results match, then concluding
18 that the two candidate search results are
19 near-duplicates, and
20 wherein cluster identifiers of the candidate
21 search results are assigned by:
22 i) determining whether or not a document
23 corresponding to the candidate search result is a
24 near-duplicate of any of previously processed
25 documents,
26 ii) if it is determined that the document
27 corresponding to the candidate search result is
28 not a near-duplicate of any previously processed
29 document, then associating the document with a
30 unique cluster identifier, and
31 iii) if it is determined that the document
32 corresponding to the candidate search result is a

X